MASSACHUSETTS INSTITUTE OF TECHNOLOGY VLSI PUBLICATIONS



VLSI Memo No. 89-518
April 1989

# A Manufacturing Scheduler's Perspective on Semiconductor Fabrication

X. Bai and S. B. Gershwin

Abstract

In this paper, we describe the phenomena in semiconductor fabrication which are important for production scheduling. We focus our attention on collecting and understanding events, and describing related concepts, such as fabrication processes, operation sets, production machines, support machines, accessories, operation worker, support technicians, activities, objectives, etc. We do not suggest scheduling policies here; instead the goal is to characterize all the events that must be considered in developing such a policy.

Microsystems Research Center Room 39-321
Massachusetts Institute of Technology
Cambridge Massachusetts 02139
Telephone (617) 253-8138

Acknowledgements

This research was supported in part by the Defense Advanced Research Projects Agency under contracts N00014-85-K-0213 and MDA972-88-K-0008.

Author Information

Bai: Laboratory for Manufacturing and Productivity, Department of Mechanical Engineering, MIT, Room 35-104, Cambridge, MA 02139. (617) 253-2730.
Gershwin: Laboratory for Manufacturing and Productivity, Department of Mechanical Engineering, MIT, Room 35-331, Cambridge, MA 02139. (617) 253-2149.

VLSI#:

LMP#: 88-004

# A Manufacturing Scheduler's Perspective On Semiconductor Fabrication

*by*


**X. Bai and S. B. Gershwin**

Laboratory for Manufacturing and Productivity

Massachusetts Institute of Technology

77 Massachusetts Avenue, Cambridge, MA 02139


March, 1989

# A Manufacturing Scheduler's Perspective On Semiconductor Fabrication

by

X. Bai and S. B. Gershwin

## Abstract

In this paper, we describe the phenomena in semiconductor fabrication which are important for production scheduling. We focus our attention on collecting and understanding events, and describing related concepts, such as fabrication processes, operation sets, production machines, support machines, accessories, operation workers, support technicians, activities, objectives, etc. We do not suggest scheduling policies here; instead the goal is to characterize all the events that must be considered in developing such a policy.

# Contents

Abstract 1

Contents 2

1 Introduction 5

2 Overview of the Semiconductor Manufacturing Procedure 6

3 A Closer Look of Semiconductor Fabrication 8
3.1 Fabrication Processes and Operation Sets . . . . . . . . . . . . . . . . 9
3.2 Inspection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
3.3 Machines . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
3.3.1 Support machines . . . . . . . . . . . . . . . . . . . . . . . . 13
3.3.2 Production machines . . . . . . . . . . . . . . . . . . . . . . 16
3.3.3 Accessories . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
3.4 Human resources involved in IC fabrication . . . . . . . . . . . . . . 19
3.4.1 Operation workers . . . . . . . . . . . . . . . . . . . . . . . . 20
3.4.2 Support technicians . . . . . . . . . . . . . . . . . . . . . . . 20
3.5 Support relationships in wafer fabrication . . . . . . . . . . . . . . . 21

4 Events and Activities in IC Fabrication 21
4.1 Controllable activities . . . . . . . . . . . . . . . . . . . . . . . . 21
4.2 Uncontrollable and predictable activities . . . . . . . . . . . . . . . . 23
4.3 Uncontrollable and unpredictable activities . . . . . . . . . . . . . . . 24

5 Constraints on the Scheduling 25

6 Scheduling Objectives and Factory Types 27
6.1 Scheduling objectives . . . . . . . . . . . . . . . . . . . . . . . . . 27
6.2 Factory types . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

7 Summary 28

8 Acknowledgment 29

References 30

# List of Figures


1 Semiconductor manufacturing procedure . . . . . . . . . . . . . . . . . 7

2 Two mask poly gate capacitor process . . . . . . . . . . . . . . . . . 10

3 Inspection mechanism . . . . . . . . . . . . . . . . . . . . . . . . . 13

4 Support relationships in wafer fabrication . . . . . . . . . . . . . . . 22

# List of Tables

1 An example of opset: dfield5k.set . . . . . . . . . . . . . . . . . . . . . . 11

# 1 Introduction

The scheduling of semiconductor fabrication is a challenging problem because of the large number of operations and very long throughput times (also called cycle times). It has attracted attention from researchers, but there is still much work to be done. Leachman [1] describes a preliminary production planning framework for semiconductor industry. Bitran and Tirupati [2] develop production planning and scheduling models for wafer production facilities. Burman *et al.* [3] analyze the performance of IC manufacturing lines using operations research techniques. Chen *et al.* [4] study a queueing network model for wafer fabrication factories. Glassey and Resende [5] and Wein [6] report the impact of job-release rules on the wafer fabrication production. These papers all focus on limited issues and use highly specialized models.

To achieve effective scheduling, we need a better understanding of the IC fabrication environment. The purpose of this work is to study the phenomena by taking a closer look at the fabrication procedure from a manufacturing scheduler's viewpoint, and to build up a knowledge base for further mathematical modeling. Since semiconductor fabrication factories can be quite different from each other in the industry, we cannot exhaustively enumerate all possible events. We try to list as many as possible of the common and important events. We do not present any scheduling models or methods here. Instead, our goal is to characterize the phenomena and events that must be treated in the development of scheduling policies.

Many of the events that we described here are random, and impossible to predict precisely or control. Consequently, we suggest that the design of scheduling techniques will be concerned with stochastic models and issues related to computational speed.

Many of the phenomena described here, such as machine failures, maintenance, and the need to satisfy customers, are common to all types of manufacturing. Others are particularly important to semiconductor fabrication, and they include low or variable yields, very long lead times, machines that can hold many wafers at the same time, large number of operations, and re-entrant flow (in which parts travel in a cycle among the same machines several times).

Terminology is not standard throughout the industry. For example, "chips", "devices", "die", or "circuits" are fabricated on "wafers" or "slices". We use the terminology of the MIT Integrated Circuits Laboratory here.

# 2 Overview of the Semiconductor Manufacturing Procedure

The overall semiconductor manufacturing procedure can be roughly divided into six subprocedures: circuit design and mask preparation, wafer preparation, wafer fabrication (or wafer fab, or IC fab), probe test and sort, assembly, test and classify. Figure 1 illustrates the manufacturing process flow for a semiconductor firm. In terms of the product structures, the six subprocedures are described in the following.

Circuit design and mask preparation: According to marketing information and technology development, new circuits are laid out with the aid of a computer. A mask is a glass plate with a hard surface material such as chromium, chromium oxide, iron oxide, or silicon. An image is created in a mask by a pattern generator which removes material using a directed electron beam in a high vacuum.

Wafer preparation: This process begins with quartz which is refined into electronics grade silicon. The silicon is grown into cylindrical crystals three to six inches in diameter. Some newer systems use eight inch wafers and there is experimental work with twelve inch wafers. The cylinders are sliced into wafers which are then buffed, polished, and possibly doped with some impurity. These wafers are then ready for fabrication of circuits.

Wafer fabrication: A wafer fabrication procedure is performed in a wafer fab factory, which contains a number of machines and a workforce. Corresponding to different final products, a number of fabrication processes are run in a wafer fab. Each of the fabrication processes is a series of processing steps the wafers pass through during the manufacturing. Each processing step is associated with an operation set(opset) which consists of several operations in series on different machines. (The opset terminology is used only in the MIT Integrated Circuit Laboratory, but we find it useful and discuss it in detail in Section 3.) Common processes include CMOS, NMOS, and Bipolar. Other processes serve to make accelerometers, flow transducers, microphones, etc. Each of the processes creates useful three-dimensional structures, such as transistors, capacitors, resistors, and transducers, on the wafers. Each potential integrated circuit device is called a die, which consists of a number of those structures.

Wafer probe and sort: Each die on a wafer is inspected by using a wafer probe. Rejects are marked (sorted) so as to be discarded in the assembly procedure. The inventory of probed wafers is called die inventory. Wafer probe process consists of

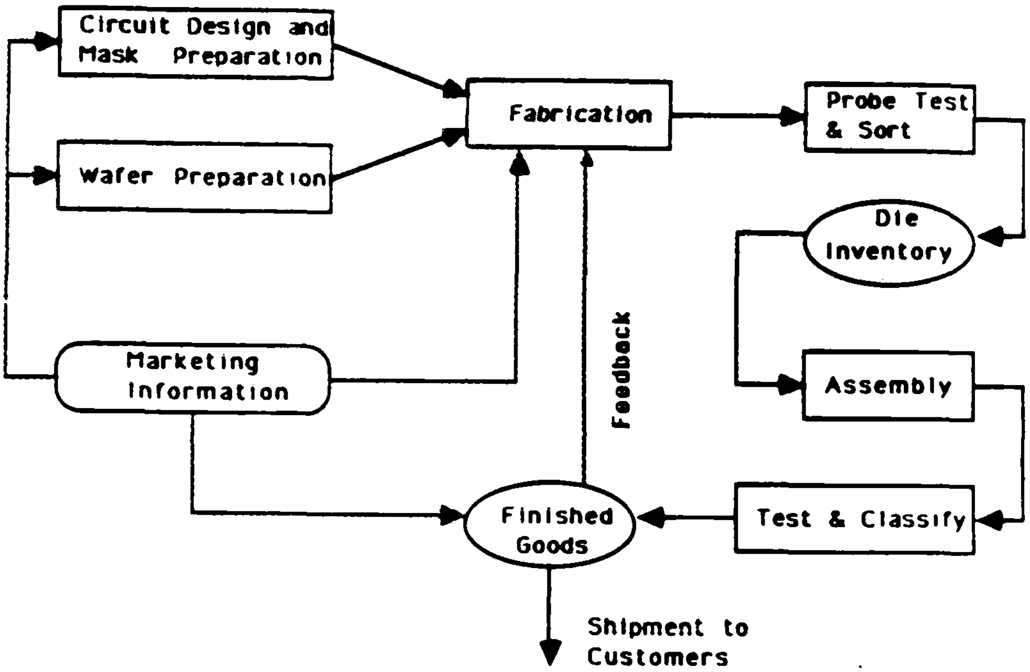


Figure 1: Semiconductor manufacturing procedure

one or only a few steps. In some firms, these steps are thought of simply as the final steps of the fabrication processes [1].

Assembly: In this process, wafers are sawed and rejected die are discarded. Good die are sealed in packages of various types [1].

Test and classify: Two test processes, raw test and final test, are involved here. In raw test (or class test) process, packaged devices are subjected to a series of tests to determine device performance. As result of this test, packaged devices are categorized into bins based on the measured performance of one or more attributes such as device speed, power consumption, tolerance of voltage variance, etc. Final tests are performed before delivering products to customers, according to the orders [1].

In this paper, we focus our attention on wafer fabrication processes to get a better understanding of the manufacturing procedure from the scheduler's viewpoint. To help describe these events, we define some terminology. A *resource* is any part of the production system that is used to perform or support fab processes. Machines, operation workers, and support technicians are resources. An *activity* is a pair of events associated with a resource. The first event corresponds to the start of the activity, and the second is the end of the activity. Only one activity can appear at a resource at any time. Operations, maintenance, and failures are activities.

In Section 3, we discuss fabrication processes, operation sets, machines, and human resources involved in IC fabrication. Activities are listed in Section 4. Constraints and objectives are discussed in Section 5 and 6, respectively.

# 3 A Closer Look of Semiconductor Fabrication

In a wafer fabrication factory, wafers are grouped in lots. They are grouped this way because many machines are designed to work on many wafers at the same time; because changing operations on some machines can be expensive; and because this makes it easier to trace the path of a wafer through the system for the purpose of determining causes of poor yield. Each lot of wafers travels from machine to machine, following a well-defined sequence of processing steps. We refer to the predefined sequence of processing steps as the fabrication process (or fab process). Different processes correspond to different product types. Each process consists of tens or even hundreds of processing steps (or unit processes) in series. Each processing step has an associated operation set (or opset) which consists of several operations in sequence and information used for the operations and processing times.

## 3.1 Fabrication Processes and Operation Sets

Usually a wafer fab factory produces more than one product. For each product type, an operation sequence is performed to create the required structures on the wafers. *An operation is a single processing function, such as pre-oxidation clean-up or photoresist coating.* The operation sequences are called fabrication processes, such as CMOS process, NMOS process, etc. Since hundreds of operations are involved in a fab process, the operations are divided into groups, called processing steps (or unit processes), according to the processing purposes.

For example, in order to transfer a pattern from a mask to the wafers, we need to coat a photoresist layer on wafer surfaces, then expose the wafers by using an aligner, and then develop them in a wet station. After inspection, the wafers are etched and stripped in a etcher, followed by another inspection. Coating, exposing and developing have the same processing purpose: to create a patterned resist layer. consequently they are grouped together as the photo step. The etching step consists of three operations, etching, stripping and inspection.

For different patterns, we need different masks for the photo step, and for different micro-structures of the wafers, we need different etching times. That means that more information is needed to distinguish the processing steps in fabrication processes, and that is the reason to introduce operation sets (or opsets). *An opset is a group of operations which form a complete clean room processing step; it contains the information of machines, operation times (see 3.3.2), and handling times (see 3.4.1); it also specifies some parameters like furnace recipe number and photomask ID.* That is, an opset contains all the information needed to perform a processing step in a fab process.

Figure 2 depicts a simple fab process for two-mask poly gate capacitor, which consists of sixteen processing steps. Each block in the graph represents a processing step. Above the dotted line in a block is the description of the processing step, and the name of the associated operation set is under the dotted line. The first processing step is field oxidation, and the associated opset is dfield5k.set, and so on. All the operations in present step must be completed before wafers go to the next step.

In the IC Laboratory of MIT, there is a baseline process, the 1.75 micron CMOS process [7], which is used to monitor equipment performance and device character-

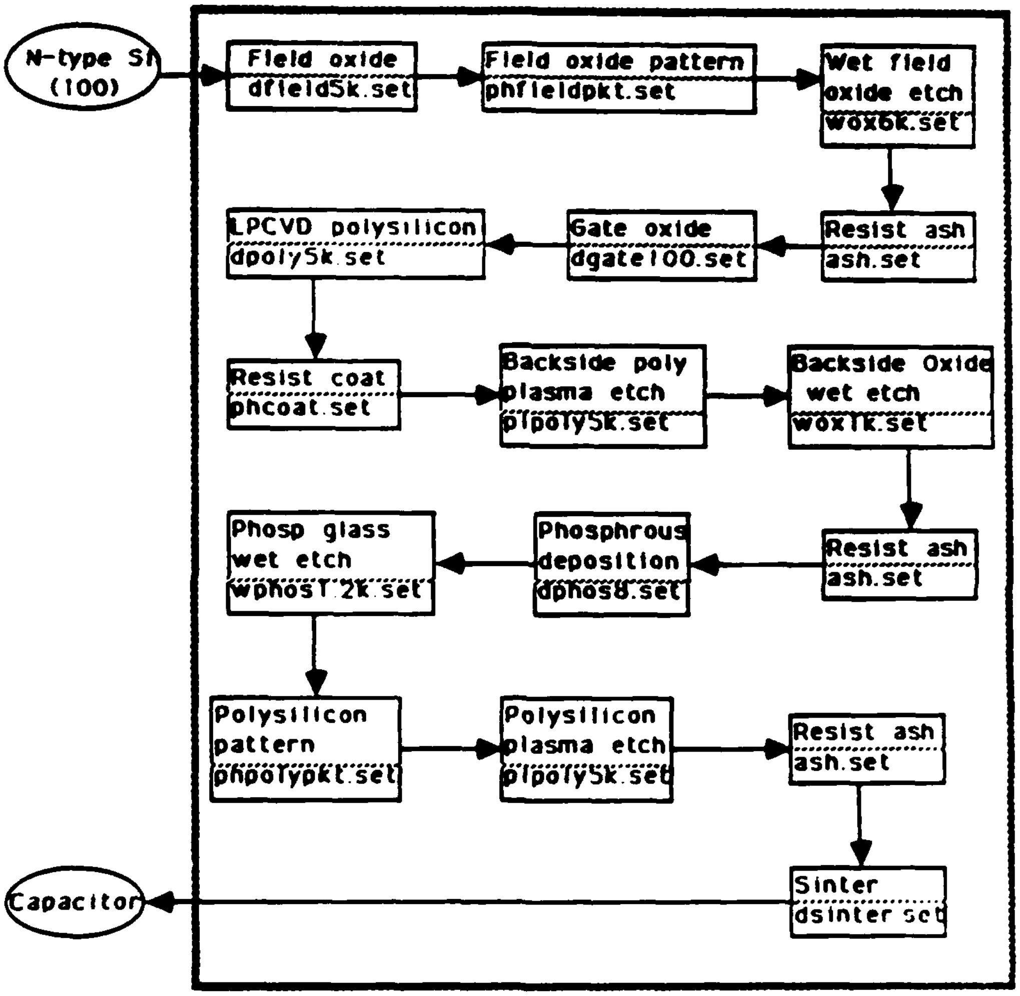


Figure 2: Two mask poly gate capacitor process

| dfield5k.set | | | | | |
|---|---|---|---|---|---|
| No. | operation | machine | parameter | op time | hdl time |
| 1 | RCA clean | RCA wet station | | 2 hrs 0 min | 2 hrs 0 min |
| 2 | diffusion | furnace B1 | recipe: 240<br>film spec<br>avg: 5100<br>range: 500 | 7 hrs 30 min | 0 hrs 30 min<br>+<br>0 hrs 30 min |
| 3 | inspection | nanospec | thickness<br>top:___<br>center:__<br>left:___<br>right:___<br>bottom:__ | 0 hrs 15 min | 0 hrs 15 min |
| | | | total time | 10 hrs 0 min | 3 hrs 30 min |

Table 1: An example of opset: dfield5k.set

istics. The baseline process is enhancement-compatible so that new technology innovations can be tested in a real integrated circuit process. The process was designed modularly such that coupling between processing steps (unit processes) was minimized. Whenever possible, these baseline processing steps should be incorporated into new processes and experiments. That means that all of other processes should be modified baseline processes.

Table 1 illutrates an example of opset, named dfield5k.set [8]. It consists of three operations, RCA clean, diffusion, and inspection, in sequence. The machines used for the operations are RCA wet station, furnace B1, and nanospec, respectively. The wafers undergoing this opset visit RCA wet station first, then to furnace B1, followed by inspection at nanospec. All the three operations must be completed before the wafers are ready for next opset. The required operation times (see 3.3.2) and handling times (see 3.4.1) are listed in the table. The total times include a 15 minute transportation time. The specified parameters provide further information to support each operation. For example, recipe# 240 contains the information about the temperature set-up and gas inputs for the furnace, and so on.

In terms of the processing purpose, opsets can be divided into small groups, such as diffusion, ion implantation, metal deposition, photo, plasma etch, wet etch, etc.

From this discussion, we see that each opset is like an independent micro-process, which completes a step of construction of the final device on the wafers, such as silicon oxide deposition, metal deposition, etc. The number of operations and the order of the operation sequence in a opset are usually fixed, and the operation times are fixed too.

*A fab process is a sequence of opsets.* A number of opsets are combined in sequence to form a process for each product type. The number of the opsets and the order of the opset sequence in a fab process may be changed to form different processes for different final products. Whenever possible the baseline processing steps should be incorporated into new processes.

There is an opset base which consists of all of the opsets involved in a wafer fab. For example, if the process in Figure 2 is the $n^{th}$ process in a wafer fab, then the associated opset of $3^{rd}$ step of the $n^{th}$ process is wox6k.set, which might be number 36 in the standard opset base.

## 3.2 Inspection

Most processing steps are ended by an inspection operation. The purpose of inspection is to control the product quality and to test machine performance. Decisions are made according to the inspection results. For example, good wafers which pass inspection are sent to downstream buffer to queue for next processing step, and bad wafers which fail inspection are either sent to upstream buffer for rework or scrapped to the trash can.

Not all of the production wafers go to inspection machines. Usually, only control (pilot) wafers or sample wafers are sent for inspection. When a wafer fails inspection, the cause of the failure is carefully searched, and then support technicians are notified to maintain or repair machines.

Figure 3 illustrates the inspection mechanism in wafer fabrication. Suppose we consider the opset (ni), associated with the $i^{th}$ step of the $n^{th}$ process, which consists of three operations,(ni1), (ni2), and (ni3). And (ni3) is an inspection operation. Wafers in lots that pass the inspection are sent to the downstream buffer $B_{ni}$ to queue for next opset (n,i+1). When wafers fail inspection, first, we need to take one of the two choices: rework or throw them away. For example, if a polysilicon layer on the wafers is wrong, we have to scrap them, but if a silicon-nitride layer is wrong, we can send the wafers to an upstream buffer for rework. Second, if we decide to rework, we have to decide which upstream buffer to send them to, according to the

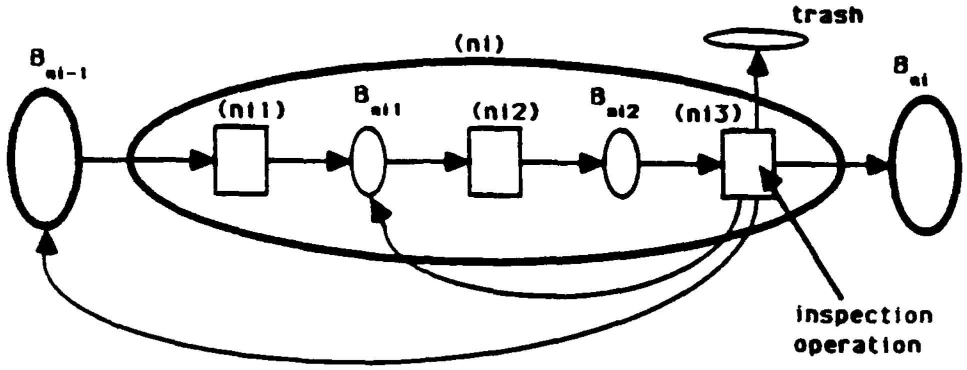


Figure 3: Inspection mechanism

inspection results. If operation (ni1) is wrong, we need to send the wafers back to the upstream buffer $B_{ni-1}$, but if only operation (ni2) failed, we should send the wafers to the inner buffer $B_{ni1}$. A knowledge base is needed to support this kind of decision making.

## 3.3 Machines

In this section we describe the machines involved in semiconductor fabrication. In terms of function of the machines, we divide them into two basic groups, i.e., support machines and production machines. We also discuss accessories. Machines in different wafer fab factories may not be exactly the same. The discussion here is based on the IC Laboratory of MIT.

### 3.3.1 Support machines

Support machines in a wafer fab are never visited by the wafers. The operation states of production machines depend on the states of these machines. If one of the support machines is down or undergoing a maintenance, one or more production machines will be down. In general, performing maintenance on the support machines causes a shut down of the clean room if the system is fully utilized. The IC Laboratory of MIT is run only twelve hours a day. Consequently, maintenance is usually done during the off time. We refer to the whole set of support machines as the house system. Following is a list of support machines in the IC Laboratory of MIT [9].

*Clean air flow*: The laminar flow of filtered air is used to maintain a dust-free environment in clean rooms. This flow is provided by fans mounted on the roof, by a maze of ducts, and by filters are located above the clean rooms. Biannually, the air system is visually inspected to check the belts, fans, and bearings. The motor is lubricated; wiring is checked; control circuits, pilot lights, and interlocks are tested. Pre filters are changed quarterly, and the bag filters are changed biannually. The service takes a few days.

*Emergency water*: This is ordinary cold city water. It is used in emergencies to wash a person splashed by chemicals. This is biannually checked for leaks and flow testing. It may take a day or two.

*Process vacuum*: This is used to pick up wafers and hold wafers in place during processing. This system is serviced only when there is a problem. It has not failed since it was installed in 1985.

*Cleaning vacuum*: These are used to vacuum the rooms clean or clean up spills. The centrally located pumps collect waste, and the tanks must be emptied when full. The hoses are attached to the plug points in the wall from which pipes lead to the machines. This is serviced only when there is a problem. It has not failed since it was installed in 1985.

*Compressed air*: This can be used to operate pneumatic equipment. A dessicant at the main pumps dries the air. Biannually, the incoming filters are changed; the pumps are serviced; belts are checked; plumbing is checked; drain traps are cleaned. The maintenance duration is a few days.

*Process chilled water (pcw)*: This is water used to control the temperature of equipment. It is recycled and reheated or recooled as required. The pump is serviced annually. Also all wiring, overloads, interlocks, control circuits, pilot lights, and voltages are checked. It is not necessary to shut down the clean room to do the maintenance. However, all related machines will be shut down when a failure occurs. This equipment was down twice at MIT during 1988, and took about one week to repair each time.

*Solvent tank*: A variety of chemicals are used and dumped. They are poured down the solvent waste sinks where they drain into a special underground storage tank and then are taken to long term storage. The storage tank will be emptied whenever it reaches 300 gallons. It is not currently used at the MIT ICL.

*City waste*: Other chemicals are neutralized to ph of 7 and dumped into the city sewer. Processing tanks are checked daily. Acid and base are added as necessary weekly. Annually, the pump is serviced, which includes lubrication and testing of wiring, voltages, and pressure differentials. The maintenance takes a few days. All wet stations are shut down when the maintenance is performed.

*Fume exhaust*: This is a system of ducts and fans that perform the work required to exhaust poisonous fumes from fume hoods. It includes a number of traps which should be removed and cleaned monthly. The fans are visually checked for belt tension quarterly. The system undergoes a full check of fans, annually, and all belts are changed. The maintenance may require a few days. The clean room will be shut down when maintenance is performed.

*Power*: Several different voltages, currents and phases are available. All of the machines require electricity to run. A backup power supply exists, so power is essentially always available. Whenever a power failure occurs, it is necessary to reset all the machines.

*Deionized (DI) water*: This can be considered as a storage tank with fixed volume and a production system with a maximum replenishment rate. It is possible to use up the DI water faster than it can be replenished. It is monitored daily. The DI lines require periodic cleaning which shuts the system down. The maintenance program is equivalent to about 2000 man hours yearly.

*Humidity controllers and temperature controllers*: Excessive moisture can damage wafers by accelerating undesirable chemical reactions on chemically active surfaces. Therefore the humidity in the clean rooms is controlled. The temperature in the clean room is controlled to within $\pm$ 1 degree because wafer dimensions depend on temperature and some equipment is highly temperature sensitive. Biannually, the bearings, dampers, and fans are lubricated. Belts are checked for tension and wear. Visually inspection of support, and vibrition check. Filter banks are also checked. Annually, motors are checked for voltage, wiring, and overloads. The maintenance may take about a week.

*Tank farm*: Semiconductor fabrication requires clean dry gases. At MIT, three tanks of Argon, Nitrogen and Oxygen supply the building with these gases and liquid nitrogen. Every day one percent of each of these tanks must be used or boiled off to keep the tanks cool and the temperature low. The tanks cannot be empty; they must stay partially filled to prevent contamination. Therefore operations which would empty a tank cannot be performed or the tank will be become contaminated. They are monitored daily and serviced as necessary by outside vendor.

*Local gases*: These are small tanks of gases placed in cabinets near the equipment which use them. They include Silane, Phosphine, Hydrogen chloride, Ammonia, Boron trichloride, Dichlorosilane, Sulfer Hexafloride, Freon 13, freon 14, freon 116, Helium, Phosphorous Pentafloride, Silicon Tetrafloride, and Boron trifloride. There are a number of gas cylinders in each cabinet. Each gas cylinder is connected through valves, to filters and then to wafer processing equipment. They are maintained annually.

*Local gas vent*: To ensure that the leakage of a gas cylinder does not poison anyone, the air from the gas cabinets is exhausted.

*Safety alarms*: Fire and gas leaks are reported by safety alarms. Fire extinquisher and fire pumps are checked annually. Hydrogen monitors and toxic gas monitors are checked daily visually and serviced monthly. Sensors are cleaned; lamps are aligned; and tapes are changed.

### 3.3.2 Production machines

Wafers visit the production machines and occupy them for certain periods of time. These production machines impose capacity constraints on the production rates. That is, no machine can be busy more than 100% of the time. The period of time required by a machine to perform an operation on the wafers is called *operation time*. Whenever we talk about an operation time, we must specify both the machine and the operation. For instance, in Table 1, furnace B1 needs 7 hours and 30 minutes to perform the diffusion operation in dfield5k.set. Some of this is load/unload time. In terms of purposes of the operations done by the machines, we name different groups of the production machines as diffusion equipment, lithographic equipment, inspection equipment, etc. Following is a list of production machines in MIT ICL.

<u>Lithographic equipment</u>:

Photolithographic operations transfer the image on a mask to the photoresist layer on a wafer. They are done in a wafer track which consists of several machines listed as follows:

*HMDS vacuum bake vapor prime and image reversal system (Model 3/10)*: Wafers are sprayed with a dehydrating chemical, HMDS, at 150° C. A dedicated commercial oven is used for this operation, which requires house vacuum, power, and a dry and particle-free environment. It is expected to fail about once every 5 months and will require about 3 days to repair.

*Photoresist coater & developer (GCA 1006 Wafertrack)*: After HMDS, wafers are loaded on the GCA wafer coating track for photoresist coating. The pre-exposure bake is done in the in-line contact oven module. After the exposure, the exposed wafers are developed on the GCA developing track. Post-development hard baking is done in a in-line oven. This equipment is expected to fail about once every 2 months and will require 3 days to repair.

*Wafer stepper system* (GCA 4800 DSW): This equipment does the exposure. The pattern transfer from an appropriate mask is carried out in a GCA 4800 , 10X direct step-on wafer system equipped with a 10-78-45 g-line lens. All the relevant information about stepping a given mask pattern on a wafer are stored in a job specification file in a dedicated PDP-11. Operation times are longer for wafers with smaller widths, due to the greater precision required for alignment. The mean-time-to-fail (MTTF) is about 2 weeks and the mean-time-to-repair (MTTR) is about 2 days.

*Asher*: In all baseline steps, resist is removed by washing in a photoresist stripper (Drytek Model Megstrip 6). The MTTF is about 2 months and the MTTR is about 3 days.

Etching equipment:

*Dry etching*: This is done in a plasma. Due to an applied voltage the ions in the plasma bombard the silicon target perpendicular to the surface. The surface is eaten away where it is not covered by resist. There are three plasma etchers in MIT ICL. Nitride etching and polysilicon etching are done in etcher-1 (LAM 480). SF6 and CCl4 are used as etching gas respectively. Etcher-2 (LAM 594) is used for oxide etching with CHF3+CF4 as etching gas. Metal etching is done in etcher-3 (LAM 690). These machines are expected to fail about once every 2 months and will require about 2 days to repair.

*Wet etching* (wet chemical process station): In addition to the dry etching steps, stripping of oxide and silicon nitride is done using wet chemistries. The wafer is placed in a bath, and the etch eats away at the parts of the layer not covered by resist. The MTTF is about 5 months and the MTTR is about 2 days.

Diffusion equipment:

*RCA clean*: Before wafers are loaded into furnaces, they are cleaned in a wet station using chemistries. It is expected to fail about once every 3 months and will require about 2 days to repair.

*Oxidation furnaces*: The MIT ICL is equipped with BTU oxidation furnaces. These machines are used to expose wafers to hot gases at a variety of pressures for oxidation or diffusion. Furnaces consist of quartz tubes, gas controllers, temperature controllers, a suspended loading system, and a dedicated PDP-11. The ICL staff, who is responsible for diffusion, maintains these furnaces with respect to cleanliness and routine chemical vapor monitoring. These machines are expected to fail about once a year and require 3 days to repair.

*Chemical vapor deposition (CVD)*: Layers can be deposited on wafers, in furnaces, from gases by a process called chemical vapor deposition (CVD). If the process occurs at low pressure, it is called low pressure chemical vapor deposition (LPCVD). There are four LPCVD tubes in MIT ICL. Deposition layers include polysilicon, silicon oxide, silicon nitride, and boro-phospho-siliocate-glass (BPSG). The MTTF is about 2 months and the MTTR is about 3 days.

Metalization equipment:

*Sputtering system* (CVC 601): Al 1%Si is the baseline first level metal. It is deposited in a CVC sputtering system in the dc megnetron sputtering mode. The MTTF is about one month and the MTTR is about 4 days.

Ion implantation equipment:

*Ion implanter*: In MIT ICL, ion implantation is done in an ETON medium current machine (model NV 3206). This machine injects ions into wafers. It is operated by a special technician and requires high voltage and high vacuum. It is expected to fail about once a month and will need about 3 days to repair.

*Rapid thermal annealer*: Ion implantation damages the crystaline structure of the wafers. After ion implantation, it might be desirable to anneal the wafers. This may be accomplished by a rapid annealer (AG 210T-02). This machine did not fail at MIT during 1986-1989.

Inspection equipment:

To ensure that the process is within tolerances and that quality is maintained, wafers are inspected by measuring certain features, such as lithographic line width, impurity concentrations, film thickness, and electrical characteristics.

*Microscope*: This is used for inspection in opsets like photo and etching. The MTTF is about 2 years and the MTTR is about 5 days.

*Surface profiler* (DEKTAK IIA): A surface profiler is essentially a phonograph needle that measures the height of the bump that it crossses. Layer thickness is measured by first carving away valleys with a lithographic process, and then measuring the height of the remaining layer. It is expected to fail about once a year and to require 5 days to repair.

*Ellipsometer* (GSC L116BL-26A): This measures the reflection of a laser beam off the measured surface. This machine did not fail during 1986-1989.

*Junction sectioner* (PIC 2015D): This machine carves a groove in the wafer, stains the wafer, and then a microscope is used to identify the depth of the dopant. This machine did not fail during 1986-1989.

*Automatic four point probe*: This measures the resistivity of the incoming silicon wafers and of the layers grown on it. The MTTF is about 1 year and the MTTR is about 5 days.

*CV-plotter* (MDC CSM-16): This machine measures the space charge capacitance as a function of reverse bias voltage on a junction. This machine did not fail during 1986-1989.

*Film thickness measurement system* (Nanospec/AFT 010-0180): This equipment measures the thickness of thin film on wafers. This equipment is expected to fail about once every 1.5 years and to require 3 days to repair.

### 3.3.3 Accessories

In a wafer fab, some consumable materials are needed to run the fab processes. For instance, clean room gowns are worn over clothing; sticky mats at the door pull dirt off shoes; beakers are needed for some operations; tools are needed for certain maintenance and activities; chemical solvents are needed for some operations; poly gloves are needed in the clean room, and so on.

Different strategies can be used for accessories in scheduling. If we assume that accessories are always available, then we can ignore these factors in mathematical models. However, accessories are only available if they have been ordered in sufficient quantity. Disruptions are possible, but their frequencies and durations depend more on how well the fab is managed than on any physical phenomena.

## 3.4 Human resources involved in IC fabrication

In a wafer fab, a workforce is needed to run the fab processes. Technicians do operations and maintain equipment. Process engineers are responsible for process design and monitoring machine performance and device characteristics. A manager is in charge of smoothing processes in a whole factory. According to responsibilities.

we group them as operation workers and support technicians.

### 3.4.1 Operation workers

This group of people is in charge of operations. They either carry wafers from one machine to another or are assigned to a certain machine and spend periods of time there to perform operations. As resources, they impose capacity constraints on production rates. That is, no person can be more than 100% busy performing operations. The period of time required by an operation worker to do an operation on a machine is called *handling time*. Depending on experience, different people need different handling times to do the same operation. We could choose the average time (or average plus one standard deviation) as the constant handling time and ask for new worker training to keep the variance of handling time as small as possible.

Whenever we talk about handling time, we should also specify the operation and the machine. For example, in Table 1, an operation worker needs 1 hour to do a diffusion operation on furnace B1, i.e., 30 minutes for loading wafers and set up temperature and 30 minutes for unloading wafers. Notice that the handling time is different from the operation time. That means between loading and unloading one machine, the worker can do operations on other machines.

In the wafer fab industry, operation workers are usually assigned to a machine. They might only know how to turn the machine on or off, load or unload wafers, and press buttons to start or end operations. In a research laboratory, operation workers carry wafers around to perform operations. They are usually more knowledgeable and more actively involved in process design and product development. In the MIT IC laboratory, about 100 graduate students are involved in wafer fabrication, and most of them are in the operation worker category. Using students as workers causes scheduling complexity because of their complex personal schedules.

### 3.4.2 Support technicians

These people do not perform operations, and so they do not impose capacity constraints on production rates directly. But they do impose capacity constraints on both of production machines and support machines. That is, no person can be more than 100% busy maintaining or repairing equipment. This group includes technicians who maintain equipment, process engineers, managers, and other non-operation workers.

### 3.5 Support relationships in wafer fabrication

Figure 4 illustrates the support relationships in wafer fabrication. In a typical wafer fabrication factory, several final product types are produced. Each product type requires a number of opsets in sequence, from the opset base, to form a fab process. The arrows from the opset base to the fab processes represent the support relations. Production machines and operation workers are called from corresponding bases to support the operations. Production machines are based on support machines, and support technicians are required by both production and support machines. It should be noticed that for a certain operation, the machine is determined in the opset, but the operation worker is not.

## 4 Events and Activities in IC Fabrication

In this section, we discuss the activities that occur in a wafer fab factory. In terms of degree of control, we categorize activities as ***controllable activities***, ***uncontrollable and predictable activities***, and ***uncontrollable and unpredictable activities***.

### 4.1 Controllable activities

This kind of activity can be arranged by a decision-maker or a manager of a wafer fab factory. Usually we can only decide when to start an activity and cannot change the time that an activity requires. But there are exceptions. For example, suppose a regular maintenance of a machine takes 3 hours for one support technician, but if two technicians work together, it may take only 2 hours. An ideal scheduling algorithm should include this kind of trade-off.

*Production operations*: The major activities in a IC fab are production operations, such as wafer processing, wafer inspection, and so on. These activities are well studied and accurately timed. Operation times are sometimes random. For instance, the probe time depends on the yield and the inspection time depends on the skill of the worker. For simplification, the operation times are usually treated as constants by using estimation.

*Set-ups*: To convert a production machine from one operation to another may require cleaning the machine, changing chemistries, or performing adjustments. All such activities are called set-up changes. Wafers cannot be processed during set-up changes, and only a restricted set of operations may be performed on the machine

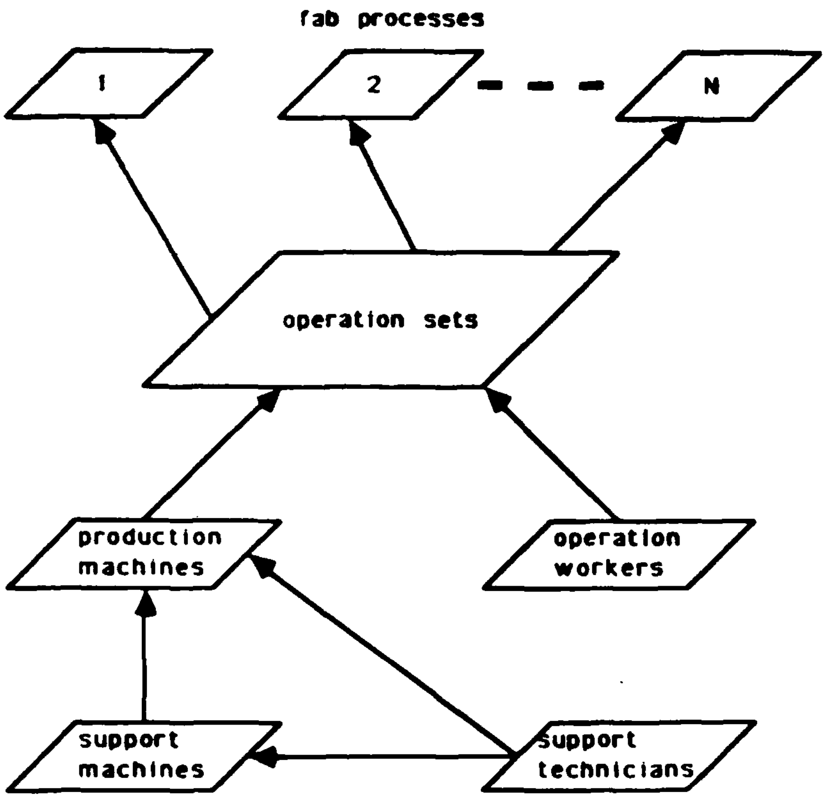


Figure 4: Support relationships in wafer fabrication

between set-up changes. Most machines, but not all, in a wafer fab factory are flexible enough to neglect the set-up change time. One of the machines for which set-up time is important is the wet-etching station for changing the chemistries.

*Multiple routings*: Sometimes, two machines can perform the same operation, and we have to decide which to use. Furthermore, one of the two machines, the primary machine, may be favored over the other due to speed or reliability. The secondary machine is sometimes used, but it may require a non-trivial change of process plan.

*Preventative maintenance*: Some regular procedures must be performed to maintain both production and support machines. If maintenance is performed late, yield may decrease, or a machine failure may occur. If maintenance is performed exactly on time it may conflict with the need to process a high priority lot. If maintenance is performed early, machine time and technician time may be wasted. Maintenance times are usually treated as constants.

*Specified events for human resources*: Both operation workers and support technicians are subject to some specified events. For example, new worker training, technician training, group meetings, overtime, and vacations are all activities that scheduling should be concerned about. We can decide where, when, and who to perform these activities, and we also can decide how long they will take.

*Environment control*: We need to do some tests regularly to ensure that yield does not suffer from variability in consumables, particle counts, humidity, machine performance, and other critical parameters. If a problem is detected, a repair may be required.

## 4.2 Uncontrollable and predictable activities

This kind of activities cannot be arranged by decision-makers or managers, but they know when they will happen and how long they will take.

*Holidays*: All human resources take holidays off. We must account for these events in advance.

*Working hours*: Some of the operation workers and support technicians are subject to special time schedules. For example, in the MIT IC fab laboratory, most of the operation workers are graduate students. Each of them has a specified class schedule. We cannot change the class schedule, but we may know when and how long they will be available.

## 4.3 Uncontrollable and unpredictable activities

For these activities, we do not know either when they will happen or how long they will take. The only thing we can do is to respond as quickly as possible. But we may have statistical data on uncontrollable and unpredictable activities for long term planning.

*Machine failure and repair*: Both production and support machines are subject to random failures and need random repair times. In general, a machine failure may be either detected during a inspection, or anticipated due to noises the machine makes prior to failing. Once a machine failure is known, the problem can be diagnosed and a repair scheduled. To some extent, we can reduce repair time by using more technicians.

*Random absence of human resources*: Both operation workers and support technicians are subject to random absences due to illness, accident, etc.

*Defective wafers*: At various points in the fab process, entire wafers are discarded, either because the wafers failed inspection or because they are broken.

*Rework*: Sometimes, one or more operations can be redone when a wafer fails inspection. There are a variety of rework policies in industry. For example, in some wafer fabs, the parent lot is held until the rework wafer catches up and then the whole lot is moved to the downstream machine. However, in other factories, the parent lot moves downstream without the rework wafer.

The rework decision on a wafer is based on some empirical rules. For instance, polysilicon deposition can not be redone while the nitride deposition can be. However, if the reason for the inspection failure of a polysilicon deposition is that the layer is thinner than the specification, the wafer can go back to the tube immediately to push the layer thicker.

*Engineering holds on lots*: Sometimes, certain process will be stopped until some experiment results are obtained or engineering decisions are made.

*Yield fluctuation*: There are two yield measurements in a wafer fab. The *process yield* is the ratio of the number of wafers which reach the probe test to the original number of wafers at the process starting point. It is random since the number of defective wafers is random. The *probe yield* is the ratio of the number of good chips to the total number of chips on a wafer. The probe yield changes randomly. The yield fluctuation affects the demand to the total number of wafers which are released to wafer fab floor.

*Product ratings*: Some devices, such as microprocessors and memory, can be delivered with different ratings for some characteristic, such as speed. These characteristics are measured at the probe test stage. A given production process may not produce devices at a single speed; rather, it produces them with a random mix of different speeds. The mix differs for different lots, and this complicates the satisfaction of customer orders [1].

*Demand change*: The production demand is a function of the customer orders and production yield and usually varies randomly. In practice, demand is often treated as constant during a short planning horizon. It is not easy to estimate the future demand since so many part types are involved in wafer fabrication. Customer cancellations are a constant hazard.

# 5 Constraints on the Scheduling

Semiconductor fabrication require machines, people, accessories, wafers, time, and so on. Each of the requirements imposes constraints on the scheduling. Here we assume that circuit design and mask preparation, wafer preparation, and process design have been done before we implement fab processes.

*Production machine capacity*: Production machines process a certain number of wafers at a time. No machine can be busy more than 100% of the time performing operations or occupied by other activities.

*Support machine availability*: Support machines do not impose capacity constraints on production rates directly. But if one support machine is down, one or more production machines may be down. That means that the availability of support machines can affect the production machine capacity.

*Down-time delay*: For some support machines, if they go down, one or more production machines will be down after a time delay. For example, if the clean air system is down, the number of dust particles will increase in the clean room. When the number reaches a certain value, all the machines which need the laminar flow of the clean air will be prevented from working.

*Operation worker capacity*: Each person takes a certain amount of time to complete an operation, and each person is limited to operations he knows how to perform. There are a limited number of people available.

*Support technician capacity*: Each technician takes a certain amount of time to maintain or repair a machine. The support technician availability affects the repair times of both production and support machines, and therefore system capacity.

*Operation sequence*: In a fab process, operations have to be performed one after another following a pre-defined sequence.

*Buffer size*: Due to technology limitations, wafers must pass through several operations before being inspected. If an operation produces defective wafers, and there is much inventory between operations, many wafers will be produced before the faulty operation is discovered. Therefore, at some points of fab processes, buffer sizes must be small. Note that the buffer size is not necessarily the physical size. Since there always is plenty of floor space to put wafers in a wafer fab factory, we might set a threshold for each buffer to control the work-in-process inventory. When the number of wafers in a buffer reaches the threshold, we stop (or block) the upstream machine.

*Limited chamber size*: The number of wafers a machine can process simultaneously is limited. For instance, the diffusion tubes in the MIT ICL can process 100 wafers in a single operation. A scheduler must sometimes decide whether to process an incomplete batch or to wait for more wafers arrive.

*Limited waiting time*: At some points in a fab process, wafers cannot wait for a long time in a buffer, because exposing the wafer surface in the air will decrease yield. For example, after RCA clean, diffusion operations must be done as soon as possible without letting the wafers wait in a buffer for a long time. Short waiting time implies small buffer size.

*Shifts*: Often wafer fab factories are run one shift per day. Wafers must be in a certain state at the end of a shift. Therefore, some operations cannot be started late in a shift, because they will not be completed during that shift.

*Pilot wafer runs*: A pilot wafer is often run through a series of processing steps before running the whole lot. By grouping compatible lots, the scheduler can increase capacity by having lots share pilot wafer runs.

*Return to same machine*: In order to increase yields, it is sometimes best to restrict a lot to a particular machine on all its visits.

*Product mix*: Sometimes, the demands for several products are in proportion. This is the case, for example, when a line is used to make the circuits for a large computer system. It is important to keep production close to the required proportions because any deviation results in useless inventory.

*The floor control system*: Many existing floor control systems in wafer fabs are old and can only handle limited amount of information. A complex algorithm may not be implementable.

*The multiple departments involved in managing a line*: In wafer fabs, more than one departments are involved in managing a production line. One department is responsible for wafer starts; another for dispatching lots within the system; another for establishing delivery schedules to customers.

# 6 Scheduling Objectives and Factory Types

In this section we discuss the objectives that scheduling of wafer fab should achieve. Usually we cannot achieve all the objectives at the same time. According to production purpose, wafer fab factories are categorized as job shop type, industry type, and mixed type.

## 6.1 Scheduling objectives

Following is a list of objectives for wafer fab scheduling.

*Inventory*: Whenever possible, we want to reduce work-in-process inventory in a wafer fab factory, because it takes up space, costs money for handling, and increases throughput time. However, too little work-in-process inventory will reduce production rates.

*Throughput time*: This is the time that a wafer spends in a fab. It is also called *cycle time* or *lead time*. The shorter the throughput time is, the faster the system can respond to customer orders, and the faster the firm can develop new products and processes.

*Variability in throughput time*: Because of the many random phenomena described here (particularly yield fluctuations) the throughput time – and therefore delivery time – can be random. Large uncertainties are undesirable.

*Market demand*: Demand is random due to market uncertainties, and due to customer cancellations. Inventory is necessary to reduce the impact of new demands, but it is a consequence of demand decreases. An important objective is to meet varying demands as close as possible with low inventory.

*Costs*: Wafer fabrication is usually very costly. For instance, some equipment depreciates at $100 per minute; and in addition to facility costs, just to maintain operating environment can cost $3 million or more per year. Whenever possible, we want to reduce the costs. Effective scheduling can reduce material costs, resource costs, energy costs, and so on.

## 6.2 Factory types

Some objectives are in conflict. For example, if we want to deliver products on time, we have to build up inventory, and if we want to reduce throughput time, we have to reduce inventory. Thus, sometimes we have to make a choice of objectives for scheduling. Based on the production purpose, categorizing factory types may help make this choice.

*Job shop type*: In some wafer fab factories, there are many product types. The production demand is usually small for each product type, and changes randomly and rapidly. For example, in a research laboratory, we only process a small number of wafers for each experiment, and the requirements are unpredictable. In a wafer fab factory, the personalization processes, which create customized circuits, involve thousands of chip part types, and highly variable demands. We categorize this kind of factory as a job shop.

*Industry type*: In this kind of factory, there are few product types. Production demand is large for each product type, and is relatively predictable.

*Mixed type*: In most wafer fabs, both the high volume and the personalization processes are performed simultaneously. Sometimes, engineers need to do experiments in a industry type factory to develop new processes and products. Two kinds of production demands are then mixed together. Such wafer fabs are categorized as mixed type.

# 7 Summary

In this paper, we have discussed many of the events associated with wafer fabrication. We have attempted to describe all the issues which are important for production scheduling. The purpose of this work has been to list and understand the events so that schedulers can be designed to account for these phenomena.

## 8 Acknowledgment

The authors are grateful to many people both inside and outside of MIT for helpful discussions and suggestions. In particular, we thank Paul Maciel, the manager of MIT ICL, Mike Schroth, the research specialist of MIT ICL, and Prof. C. G. Sodini for their thoughtful suggestions and comments. We would like to thank C. Lozinski for providing us the unpublished material [9]. We thank L. P. Devaney, the research specialist of MIT ICL, for providing us the preventative maintenance information about the house system of MIT ICL. We also thank S. Hood and D. Connors of IBM, J. Hogge and Y. Choong of Texas Instruments, and W. J. Trybula of General Electronics for their comments and suggestions. This work was supported by the Defense Advanced Research Projects Agency and monitored by ONR under contract N00014-85-k-0213.

# References

[1] R. C. Leachman, "Preliminary Design and Development of A Corporate-Level Production Planning System for the Semiconductor Industry," Tech. Rep., ORC#:86-11, University of California at Berkeley, 1986.

[2] G. R. Bitran and D. Tirupati, "Planning and Scheduling for Epitaxial Wafer Production Facilities," *Operations Research, Vol. 36, No. 1*, 1988.

[3] D. Y. Burman, F. J. Gurrola-Gal, A. Nozari, S. Sathaye, and J. P. Sitarik, "Performance Analysis Techniques for IC Manufacturing Lines," *AT&T Technical Journal, Vol. 65, No. 4*, 1986.

[4] H. Chen, M. Harrison, A. Mandelbaum, A. Van Ackere, and L. Wein, "Empirical Evaluation of A Queueing Network Model for Semicondctuor Wafer Fabrication," *Operations Research, Vol. 36, No. 2*, 1988.

[5] C. R. Glassey and M. G. C. Resende, "Close-Loop Job Release for VLSI Circuit Manufacturing," Tech. Rep., ORC#:87-8a, University of California at Berkeley, 1986.

[6] L. M. Wein, "Scheduling Semiconductor Wafer Fabrication," 1987. Unpublished manuscript, Department of Operations Research, Stanford University.

[7] P. K. Tedrow and C. G. Sodini, "MIT Twin-Well CMOS Process," 1988. Unpulished MIT ICL document.

[8] "Operation Sets," 1987. Unpublished MIT ICL document.

[9] C. Lozinski, "A Scheduling Perspective on Semiconductor Wafer Fabrication," 1987. Unpublished manuscript, University of California at Berkeley.